WATER QUALITY SECTION

REGRESSION TECHNIQUES

for

ANALYTICAL CHEMISTRY TECHNICIANS

Michael W. Rawlings
Water Quality Section
January, 1991

WATER QUALITY SECTION

REGRESSION TECHNIQUES

for

ANALYTICAL CHEMISTRY TECHNICIANS

Michael W. Rawlings
Water Quality Section
January, 1991

REGRESSION TECHNIQUES

for ANALYTICAL CHEMISTRY TECHNICIANS

January, 1991...

1. A Surprise about the "Linear" in Linear Regression

Most of us unnecessarily limit linear regression. We misapply the word, "linear". Linear regression is not confined to straight lines of the form $y = \beta_0 + \beta_1 x$. The "linear" refers to the coefficients $\beta_0, \beta_1, \ldots \beta_n$, and not to the independent variables. The following functional forms, many describing curved lines, can all fitted using linear regression:

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$y = \beta_0 + \beta_1 \ln(x)$$

$$y = \beta_0 + \beta_1 (1/\sqrt{x})$$

$$y = \beta_0 + \beta_1 w + \beta_2 x + \beta_3 z$$

The first four of these are *univariate,* because the variation in the y value depends on the variation of a single independent variable, x. The complexity of the dependence is not important. The fifth equation shows a *multivariate* relationship, with three independent variables.

Fitting chromatographic peak shapes to a standard form for multiple peak resolution is a well known application of non-linear regression. The Gaussian equation is used as the peak shape model:

$$y = \beta_1 \exp[-(x-\beta_2)^2/\beta_3^2]$$

This equation is non-linear in β_2 and β_3.

We will not deal with non-linear regression. The math of the linear cases is a little formidable for most non-mathematicians; the non-linear cases can be positively esoteric. The linear regression and matrix calculations offered in LOTUS 1-2-3 can handle the tedious calculations of the linear cases for us, except when the magnitude of the numbers gets out of hand. In practical cases, that is rare.

Now that we know we need not fear the mechanical aspect of the calculations, we'll review several advanced aspects of linear regression descriptively, including some underlying theory. The object is to understand when in our daily analytical work linear regression is applicable, what form we should use, and why. We will also learn what errors arise when we ignore failures to comply with assumptions implicit in the underlying theory.

2. Rules of the Linear Regression Game

We first assume the values of the independent variables (x-axis) are known without error. All error in the location of a point lies in the y-axis direction. In our work, the x-axis values often apply to pure standards or reference materials, whose concentrations are well known and which can be prepared with accuracy. Calibration applications and analytical techniques such as standard addition satisfy this rule.

The second assumption is that the errors are normally distributed (see Appendix 1). This rule applies to almost all of the statistical procedures commonly used by technicians. Microbiological procedures produce data which is commonly treated for log-normally distributed errors. Counting procedures, e.g., radioactivity measurements, produce data where the errors conform to a Poisson distribution. Routine analytical data often appears to be contaminated with outliers. We will see how to handle these obstacles.

The third assumption is that the standard deviation associated with each data point is uniform. (For trivia nuts, this condition is called homoscedasticity.) This assumption falls out of the choice of least squares as a criterion for selecting the line of best fit. An example will help. The regression residual of each point is the vertical distance from the point to the regression line. The least squares criterion positions the line so that the sum of the squares of the residuals is a minimum. Only the magnitude of the residual is taken into account. The precision of each point is not considered in determining the position of the line.

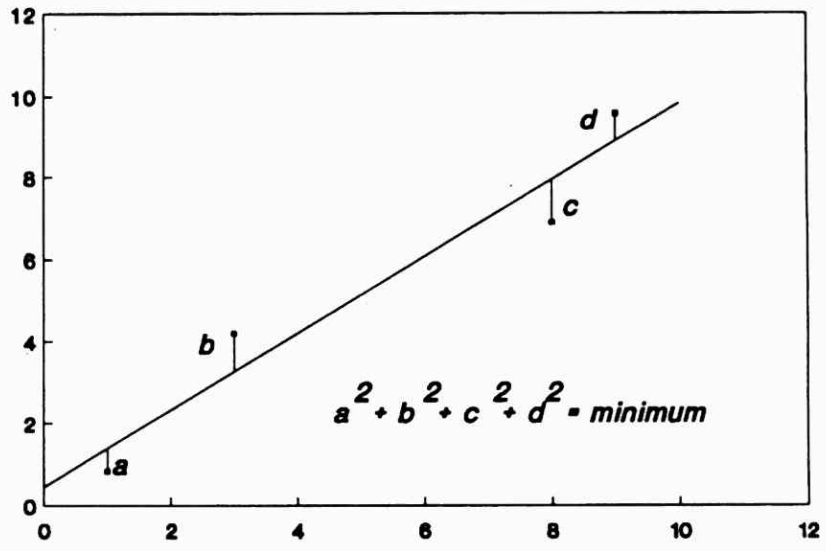


Figure 1. The Least Squares Rationale

That the standard deviation is unimportant must therefore be one element of the rationale for choosing least squares as the criterion for deciding the position of the line of best fit. This makes sense if and only if the

standard deviation is uniform over the entire range of the dependent variable. Fortunately, there are manageable regression methods which allow the standard deviation of the points to be factored in. We will examine these shortly.

Summarizing these three basic assumptions, then:

1. - the independent variable values are known without error;

2. - the dependent variable errors are normally distributed;

3. - the standard deviation of the dependent variable values is the same for all.

3. When the Rules Are Broken

Frequently we will find that at least one of the assumptions of ordinary linear regression is breached. We shall now cease to turn a blind eye towards these infractions. We have the tools at hand to deal properly with the added complexity.

3.1 Independent Variable is Not Known Exactly

This rule is clearly violated when ordinary regression is applied to method intercomparison studies. The x-axis errors are usually on a par with the y-axis errors. Ordinary linear regression will give a slope that is too steep, an intercept which is too low, and will underestimate the standard errors associated with the coefficients. Confidence limits on the line will be too optimistic, giving the misleading impression that the agreement of the methods is better than it actually is.

We can determine the first order line of best fit $y = \beta_0 + \beta_1 x$ using a maximum liklihood functional relationship approach. This approach accounts for the standard deviations of the data on both axes. Consult our *Standard Operating Procedure for Method Intercomparison* (4), for a detailed description and how-to. The author has not investigated functional relationships other than the form $y = \beta_0 + \beta_1 x$. Method intercomparison studies are, however, the almost exclusive application we have for the MLFR technique, and the first-order relationship is the likeliest model.

3.2 Non-normal Distribution of Errors

Failures to meet the distributional criterion are most often known in advance, from theoretical considerations of the analytical techniques.

Microbiologists almost automatically take logarithms of their counts to normalize the data before beginning statistical analysis. Instrumental counting methods, such as for radioactive isotopes, are best treated by Poisson distribution (see Appendix 2) statistics. There will be cases, however, where a transform or alternate statistics are unavailable. Handle these cases by taking several readings of each sample, calculating the means, and doing the regression on the means. The error distribution of means is always normal, regardless of the distribution in the original data.

Errors in routine analytical chemistry data generally follow the normal pattern with respect to central tendency and symmetry. However, the distribution is frequently long-tailed. Outliers, or what we perceive to be outliers, often haunt our data sets. The standard statistical treatment is to use an outlier rejection procedure, such as Dixon's Q test, to identify and eliminate them. These procedures presume normally distributed errors.

Outlier tests are weak and unsatisfactory from three viewpoints. The first is that many of them will fail if there is more than one outlier on the same side of the data distribution. The second is that many experts in the field of statistics feel that the distributions of scientific measurements are notably broader than normal. Thirdly, a ruthless application of common sense dictates that you will not discard any data simply because it fails to fit a presumed distributional model. If you cannot find a good reason, such as an operator blunder or a clear malfunction of equipment, to impugn a data point, keep it. Otherwise, you may seriously underestimate the standard deviation of your measurement process.

An extreme value in a small data set can be overly influential on the standard deviation, and catastrophic for the mean. We need a method allowing us to keep the extreme values, but frustrate their bad behaviour. I have lifted the following how-to section on a robust (i.e. outlier resistant) method of calculating the mean and standard deviation from the *Standard Operating Procedure for Method Intercomparison* (4). I can also supply computer programs implementing the calculation. Use them.

Suppose that we have a set of replicate sample measurements X_i, which may include outliers. "Winsorize" the data set by applying the function:

$$\varpi_i = \begin{cases} X_i & \text{if } |X_i - m| <= cs \\ (m - cs) & \text{if } X_i < (m - cs) \\ (m + cs) & \text{if } X_i > (m + cs) \end{cases}$$

where m is the mean and s is the standard deviation of the Winsorized values, and c is a factor limiting the allowable spread of data.

Calculate the mean as the simple average,

$$m = \Sigma\varpi_i/n.$$

The standard deviation of the Winsorized values is the ordinary variance modified by a factor, β, which corrects for the cut-off of the distribution's tails by the Winsorization process.

$$s = \sqrt{[var(\varpi_i)/\beta]}$$

The factor β is related to c, the factor limiting the spread of the Winsorized data set. A value of 1.5 is widely accepted and is recommended for c; the corresponding value of β is 0.778. These values protect against about 5% of outliers in a data set. The value of c should be modified to

$$c = c\sqrt{(1-1/n)}$$

to adjust the value for statistically small data sets. The value of β remains unchanged.

The calculation of the pseudo-data set of Winsorized values requires the set's own mean and standard deviation. Apparently, you have to know the answer *before* you do the calculation. We use an iterative process to resolve this chicken-or-egg situation. Use the estimates of m and s from the current calculated set as "seed" values for the next set. Repeat this until convergence is achieved. Suitable starting values for m and s are:

$$m_0 = \text{median}(X_i)$$
$$s_0 = \text{median}(|X_i - m_0|)/[0.6745\sqrt{(1-1/n)}]$$

Convergence to the point that no change occurs in the fourth or fifth figure of m and s on successive iterations is adequate. This is quickly reached by a program running on a personal computer.

Note that the values m and s are *not* the mean and standard deviation of the observations. This is an advantage, not a liability. The errors in the data may not be normally distributed, and outliers may be present. The values m and s represent estimates based on the reliable part of the data, and presumably are better estimates of the true (population) mean and standard deviation than we would otherwise calculate. Bear in mind that the customary mean and standard deviation are themselves only estimates. In opting for the robust procedure, we are not sacrificing what we think of as true values for estimates - we are replacing a traditional estimate with a better one. And we retain all of our hard-won data in the process.

3.3 Non-Uniform Variance in the Data

Ordinary linear regression presumes that the standard deviation of the data remains constant over its range. Look at the Duplicates Data tables in any of the Water Quality Section Performance Reports. I have yet to find one procedure that meets the uniform variance criterion.

If I were to state, "Quantitative analysis data is always heteroscedastic.", I would rarely be incorrect. Bluntly, using ordinary linear regression even for very well behaved data such as calibration curves, is wrong. When we do so, the points at the high end of the curve - with usually far greater error associated with them - have as much or more influence on the position of the line than the low points. The net result is that the line will almost certainly be wrongly positioned near the most precisely known points. The most serious effect, on calibration curves for instance, is to mis-locate the intercept. That gives measurably false values for blanks and low level samples.

We can show off our scientific finesse by applying weighted linear regression techniques to our data. These incorporate the *y*-axis variances into the process. With personal computers to do the tedious mechanical arithmetic, we have no justification for not learning how.

4. Linear Regression with LOTUS 1-2-3

Begin by reviewing the section on / **Data Regression** in the LOTUS manual, if you're new to it or out of practice. The first order regression case is almost trivial: x values in one column, y values in the next, define your ranges, and execute the function.

For non-first order cases, a little guidance is in order. Head one column each for the functions of x appearing in your model, and a column for y. The columns for the x data must be adjacent. For instance, if you're fitting the model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^3 + \beta_3 / \sqrt{x}$$

your column headings will be:

x	x^3	1/sqrt(x)	y

Enter the x and y data values in their respective columns. In the first row of data, under each of the x-function columns, enter the formula indicated by the function to calculate the numerical value from the x column. In our

example, if the x values start in cell A5, cells B5 and C5 will contain:

+A5^3 1/@SQRT(A5)

Copy the formulae down the worksheet to cover all the x values. Check the results to be sure there were no x values which produced illegal operations, e.g division by zero or square root of a negative number.

Select **/DRX** from the LOTUS menu. Highlight the three columns containing the x data in a block as the X range of data. Select **Y**, and highlight the column of y values as the Y range of data. Select **Output range**, and highlight the cell marking the upper left corner of the range where you want your regression results to appear. In this case, the range will be five columns wide and nine rows deep. Select **Go** to calculate the regression and display the results. The output display is self-explanatory. Note that the order of the coefficients of the functions of x follow the order of the columns in the x data block.

Table 1 shows a worksheet for the example given above, with some data and the answers. Try setting this up, entering the x and y data, and entering the formulae for the functions of x. It should take about five minutes. Try your graphing skills as well; plot up the data. You should get a curve matching the figure accompanying Table 1.

Approach multivariate linear regressions in exactly the same way. Head adjacent columns for each of the independent variables, and enter the values rather than use formulae. In fact, you can mix in columns with formulae if any of your independent variables are a function of some of the others. This block of independent variables is your "X" range. You know the rest.

5. Weighted Linear Regression: the First Order Case

Variance is the ordinary measure of precision. We usually express it as the square root and call it standard deviation, because the numbers are more meaningful to us. For any two data points, the one with the lower variance is the more precise, and it should carry more weight in the regression calculations. We accomplish the weighting by scaling each point by the reciprocal of its variance.

Two points to note. First, the absolute values of the variances are not predictable from theory. Intuitively, then, it is the relative values of the variances which is important in deriving the weights. We can, and will, scale them further to make the calculations easier. Secondly, and this is so

obvious it hurts, you have to know the variances. The time to think about that is back when you're setting up your experiments, not after the fact. For routine calibration curves, you could set up a table or an equation, based on a one-time only experiment and the notion that variance varies smoothly with concentration, which provides estimates of the variances. For data from samples run on routine tests, you may be able to predict the variances from the routine duplicates data. In other cases, you must be sure to run enough replicates to get a solid estimate of the variance for each sample.

When you estimate the variance by means of multiple determinations on each sample, use the standard error of the mean instead of the standard deviation as the basis for the weighting factor. The factor can reflect a varying number of determinations of each sample, as well as the precision. The standard error of the mean of n trials is given by $s/\sqrt{n}$.

LOTUS 1-2-3 provides no function for weighted regression. It does, however, make the calculations a snap. Miller and Miller (1) give an excellent description of the calculations in section 5.10 of their book. I will not repeat their discussion; the method follows for your convenience.

Set up worksheet columns as follows; the x, y, and s (s being the standard deviation or standard error) values are entries, while the remaining columns and sums are calculations.

$$x_i \quad y_i \quad s_i \quad 1/s_i^2 \quad w_i \quad w_ix_i \quad w_iy_i \quad w_ix_iy_i \quad w_ix_i^2$$

Calculate the sums of the last six columns using the @SUM(....) function. The weighting factors, w_i, are given by:

$$w_i = s_i^{-2}/(\Sigma s_i^{-2}/n)$$

Note that the weighting factors are simply the weights divided by the average weight (use the @AVG(...) function). This results in the sum of the factors being equal to the number of points, and simplifies the calculations.

Beware: do not confuse the weights with the weighting factors. We use the weighting factors in determining the regression coefficients. We will later evaluate the results of the regression process; there, we will use the weights.

You will also need the weighted centroids,

$$\bar{x} = \Sigma w_ix_i/n \quad \text{and} \quad \bar{y} = \Sigma w_iy_i/n$$

Obtain the slope and intercept from the equations:

$$\beta_1 = \frac{\Sigma w_i x_i y_i - n\bar{x}\bar{y}}{\Sigma w_i x_i^2 - n\bar{x}^2}$$

and

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Miller and Miller provide the following example data:

x_i : 0, 2, 4, 6, 8, 10
y_i : 0.009, 0.158, 0.301, 0.472, 0.577, 0.739
s_i : 0.001, 0.004, 0.010, 0.013, 0.017, 0.022

from which the results are: $\beta_0 = 0.0091$ and $\beta_1 = 0.0738$

Model your set-up after the presentation in Table 2, and test it using this data. Consult the textbook if you have trouble obtaining correct answers. The text also shows how to calculate confidence limits (more about them, shortly) on a concentration calculated from the equation. You may borrow a copy of the text from any of the Water Quality Section supervisors.

I have shown the ordinary linear regression results for the test data for comparison, along with graphs of both functions. One should not expect to see a great deal of difference in the location of the line of best fit according to the two methods, because this data is reasonably precise. (It's a lot like our calibration data.) Remember, though, why you're doing the regression in the first place. You intend to use the relationship you discover to make further y-value predictions, or to calibrate your instrument (x-value predictions from measured y-values).

6. Assessing the Goodness of Fit: Introducing the χ^2 Test

One always wants to know how well the data fits the line, i.e., how good our model is. The temptation is to fall back on the familiar correlation coefficient, r, particularly for straight line fits. You will often see statements in the literature saying that since the correlation coefficient was, e.g., 0.993 for a linear regression, a linear relationship is confirmed. The correlation coefficient supports no such conclusion (5).

When Pearson introduced the correlation coefficient in 1903, he demonstrated that it tested whether *any* relationship existed between two sets of data. It is only happenstance that perfectly linear data yields a

value of $r = 1$. Truly curvilinear data sets often yield a correlation coefficient of 0.99 and greater, especially in applications such as calibration. The literature also demonstrates cases where the statistic fails to recognize a visibly obvious relationship.

For each point x in our data set, we can calculate a predicted value of y and compare it with the measured value. We can then determine whether these differences, the residuals, exceed those which may be expected due to chance alone, and thereby prove our model. The Chi-squared (χ^2) distribution is the basis of this test, logically enough called the χ^2 test.

The test statistic is the weighted residual sum-of-squares, W. If you are using the standard error of the means of n replicates to obtain the weights for m data points, then:

$$W = \Sigma n_i(y_i - \alpha - \beta x_i)^2/s_i^2$$

which has approximately a χ^2 distribution with m-2 degrees of freedom. If you have estimated the weights from prior knowledge of the standard deviation, use:

$$W = \Sigma(y_i - \alpha - \beta x_i)^2/s^2$$

which has exactly a χ^2 distribution with n-2 degrees of freedom. Compare the calculated statistic with the tabulated χ^2 value for the correct degrees of freedom and the desired confidence level, usually 95%. If your test value is less than the tabulated value, you may conclude that the model you have chosen properly represents the data.

Note that in both the above equations, we are summing the squared residual of each point (the quantity in the brackets) times its weight. The weights are n/s^2 or $1/s^2$, whichever is appropriate to our data. They come directly from our data set-up table, column D in the example.

You may have such a number of points that the degrees of freedom exceeds the bounds of your χ^2 table. Whitney (6) provides handy functions for calculating χ^2 values for large n and all common confidence levels. The most useful are:

$$\chi^2 = (1.64 + \sqrt{(2f-1)})^2/2 \qquad \text{for } \alpha=0.05 \text{ (95\% confidence)}$$

and

$$\chi^2 = (2.33 + \sqrt{(2f-1)})^2/2 \qquad \text{for } \alpha=0.01 \text{ (99\% confidence)}$$

where f is the degrees of freedom.

7. Confidence Limits, and How They Relate to Regression

You will rarely, if ever, have to calculate confidence limits for your regressions. If you do, Miller and Miller (1) give excellent coverage of the subject, with worked examples. But do a little empirical thinking about confidence limits with me, anyway. It will help you assess the worth of your calibration curves and QC results, and will make you more aware of the behaviour of some of the errors inherent in your system.

We have acknowledged, by using regression in the first place, that there is error in the y-values. By using weighted regression, we have further acknowledged that this uncertainty varies with the independent variable. Thus there will be an error associated with each predicted value, the magnitude of which is governed in part by the uncertainty in the location of the regression curve at the point of prediction. We would like to define limits which we are sure, to a stated degree of certainty, will enclose the true line relating the variables. The usually quoted degree of certainty is 95%, or $\alpha = 0.05$. We can calculate a pair of lines (see the figures with Table 2) which define confidence limits, or more descriptively a confidence band, bracketing the regression line. Whether we predict y-values from given x-values, or estimate x-values from measured y-values, this band defines the limits within which we expect the true value to be found 95% of the time.

The interesting thing about confidence bands is their shape. The shape is difficult to see in our ordinary regression example because the test data is so precise (Miller and Miller show an exaggerated picture). It is bow-tie shaped, with the narrowest point located at the point $y = \bar{y}$ and $x = \bar{x}$. The symmetry of the shape arises because the standard deviation is assumed to be uniform along the line of best fit. The flaring at the ends happens because as you move closer to the ends, you have fewer and fewer data points on your outboard side with which to confirm your positioning. Thus you are less certain about whether the line adequately describes the location of the points; the limits widen accordingly.

The same flaring is present in the weighted regression case, but added to that is the effect of the known standard deviation of the y-values. In our test example, the standard deviation varies by a factor of 20 over the range of the values, and is lowest at the low end. We account for the variation in the standard deviation in weighted regression, rather than leaving the regression mechanism to assume it. The constriction of the confidence band about the more precise values is the mathematical manifestation of that accounting.

A word on the extent of the usefulness of all this. The errors we are talking about are only those associated with the process of locating the line of best fit, and with predicting values in reference to that line. We have not, and cannot by this process alone, assess errors in the steps performed in order to produce the readings which make up our data. The value in defining the errors in the regression is that one may factor them out of a total error assessment, leaving other sources of error exposed. You may then examine and minimize the sources of error in the bench work. When you do that, you must be aware that the proportioning of error between the bench sources and the mathematical (statistical) sources changes with concentration, and by how much. The knowledge may save you a lot of time chasing up blind alleys, experimentally.

8. On Preparing for Multivariate Weighted Linear Regression

Most of our continuous flow analyzer systems are calibrated assuming a parabolic, or second-order, curve. I have not been able as yet to work out or find references giving a summations solution for weighted parabolic regression. It seems that the summation expressions become unmanageable in size and complexity. The alternative is to return to the basics of the math of linear regression: matrix procedures. All varieties of linear regression, from the simplest to the most complex, are handled identically in matrix notation.

We're poorly schooled in matrix math. There's reason for that; it's a very tedious, error prone process to do by hand, so it has been ignored. Nowadays it's not difficult at all with assistants like LOTUS 1-2-3 on call. Skip ahead a little and look at the spreadsheet display in Table 3. Most of the numbers you see are the results of operations easily set up by you and performed by LOTUS. You will enter no more values than you did for the weighted first order case, and fewer formulae. Be assured that perfectly lucid instructions for doing these calculations on any data set follow in a little while. In the meantime, we can profitably learn a little bit of matrix notation, and something of the operations we will set LOTUS to work upon.

9. Some Basic Definitions and Conventions of Notation

Whenever you have made a list of numbers, such as concentrations of standards together with their responses, you have set up a matrix. A **matrix** is simply an ordered, rectangular array of numbers, arranged in rows and columns. A matrix is described by its **order**, which is the number rows followed by the number of columns. For example, a 2x3

matrix is an array of two rows of three columns. Specifying the rows first and columns second is conventional, and is known as **row-major** order. Each individual entry in the matrix is called an **element**. The position of any element in an array is shown by listing the algebraic symbol for the element subscripted by the row and column numbers, as follows:

$$a_{2,4} \text{ or } a_{i,j}$$

In the first example, the element resides in the second row, fourth column. The second example is the general case, where $_i$ is the row number running from 1 to the maximum number of rows, and $_j$ is the column number running from 1 to the maximum number of columns. When you write out a matrix in array form, enclose it in square brackets.

There are three special matrices which are important in regression. A 1x1 matrix consists of a single element only, and is known as a **scalar**. A 1xn or an nx1 matrix, which consists of a single row or column respectively, is called a **vector**. A one-row matrix is a **row-vector**, and a one-column matrix is a **column-vector**. A **diagonal** matrix can be of any square order (n,n), but has non-zero elements only on the diagonal (positions 1,1 to n,n) and 0 elsewhere. Freund and Minton (2) describe other special matrices.

The notation for matrices varies somewhat from text to text. Most, however, are quite clear, and one of the best systems is the one we will use here.

A **scalar** is represented by an ordinary algebraic symbol, in italics: *z*. If it is a constant, it may be replaced by its numeric value.

A **vector** is represented by a bold lower case letter: **e**.

A **matrix** is designated by a bold upper case letter: **A**.

The **elements** of a matrix are designated by the subscripted lower case of the letter which designates the matrix: $m_{i,j}$ are elements of **M**.

The **identity** matrix is always **I**. This is a diagonal matrix with all diagonal elements equal to 1.

10. Some Basic Matrix Operations

You treat matrices much like ordinary numbers, using many of the same operations. You must observe some requirements of structure and order of operations, but the operational concepts are familiar.

Transposition. You can transpose any order of matrix. The operation is equivalent to interchanging the row and column indices of every element:

$$m_{i,j} \rightarrow m_{j,i}$$

In notation you will see: $\mathbf{M'} + \mathbf{A} = \mathbf{B}$

Transposition is not usually called for as a separate operation; the primed notation appearing in a calculation indicates you are to use the transposed matrix. The example directs us to add the transpose of **M** to **A**. The easiest way to transpose by hand is to read down the first column, and copy the elements in order into the first row. Repeat this for each subsequent column and row in the matrix.

Addition and Subtraction. You may add or subtract any two matrices of the same order. If you have two matrices **A** and **B**, and wish to form the summation matrix **C**, then:

$$a_{i,j} + b_{i,j} = c_{i,j} \quad \text{for all i,j}$$

Symbolically: $\mathbf{A} + \mathbf{B} = \mathbf{C}$

Multiplication. You may multiply any two matrices where the number of columns in the multiplicand is equal to the number of rows in the multiplier. The product matrix will have the same number of rows as the multiplicand and the same number of columns as the multiplier. Freund and Minton (2) show the multiplication procedure in detail, if you're interested. We will let LOTUS do the calculations.

Symbolically: $\mathbf{A}\cdot\mathbf{B} = \mathbf{C}$

Some general properties of matrix multiplication are important with respect to the order of operations. In general, $\mathbf{A}\cdot\mathbf{B} \neq \mathbf{B}\cdot\mathbf{A}$. Also, $(\mathbf{A}\cdot\mathbf{B})' = \mathbf{B}'\cdot\mathbf{A}'$. For serial operations, $(\mathbf{A}\cdot\mathbf{B})\cdot\mathbf{C} = \mathbf{A}\cdot(\mathbf{B}\cdot\mathbf{C})$, and $\mathbf{A}\cdot(\mathbf{B} + \mathbf{C}) = \mathbf{A}\cdot\mathbf{B} + \mathbf{A}\cdot\mathbf{C}$.

There is also a separate operation known as scalar multiplication, where each element of a matrix is multiplied by a scalar:

$$z\cdot\mathbf{A} = z\cdot a_{i,j} \quad \text{for all i,j}$$

Inversion. There is no matrix operation equivalent to division. The analogous operation is matrix inversion. Difficult to do or describe, It is best known for the property that:

$$\mathbf{A}\cdot\mathbf{A}^{-1} = \mathbf{I}$$

Like transposition, one indicates that the inverse of a matrix is to be used in a calculation by using the -1 superscript. Some matrices, designated as **singular**, do not have an inverse. You may not know you have one until you invoke the inversion operation. You can probably guess that LOTUS has the inversion function built in. The HELP screen which is available if the function miscarries explains several possible causes and cures, should you ever hit a singular matrix. Freund and Minton (2) go through a tacky manual procedure to invert a matrix, for the interested.

11. The Matrix Solution for Weighted Linear Regression

Those who do not care to root around in a bit of math can skip on to Section 12. There we will implement in LOTUS 1-2-3 what we are going to develop here in matrix notation: the solution for weighted linear regression. The solution we are going to go through, however, is an application of the general solution for a set of simultaneous linear equations. It may be of interest to some for other applications.

We will define the model we wish to treat by regression as a polynomial in x, since our real example is a parabola ($y = \beta_0 + \beta_1 x + \beta_2 x^2$). We could easily replace x^2 with z, say, and x^3 with k, etc., and have a polynomial in multiple independent variables. Functions of the variables can also be accepted. Linear combinations, such as $\beta_n x + 2\beta_m x$, are forbidden because they give rise to singular matrices. In these cases, simply re-write the equations combining the related terms.

Let's use a calibration curve as an example. For that curve, we would have prepared a series of n standards of concentrations $x_1, x_2, \ldots x_n$. For each of these we have obtained a measured response $y_1, y_2, \ldots y_n$. Each measured response is made up of the true response plus some error:

$$y_{\text{MEASURED}} = y_{\text{TRUE}} + e \qquad (1)$$

If there were no error, our polynomial would be

$$y_{\text{TRUE}} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots\ldots + \beta_m x^m \qquad (2)$$

where m is the highest power of x needed to fit the curve. Recall also, that we have assumed that the x-values are error free, i.e. they are true.

Substituting (2) back into (1), we can write a separate equation for each of our standards, creating a set of n equations in n unknowns:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots\ldots + \beta_m x_1^m + e_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \ldots\ldots + \beta_m x_2^m + e_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \ldots\ldots + \beta_m x_n^m + e_n \end{aligned}$$

The n unknowns are the errors. We know the x and y values, and we are going to "try" known values for the β's until the errors are minimized.

There are three vectors and one matrix that we can separate out of this array of equations. These are the measurement vector in y, the error vector in e, and the coefficient vector in β:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_m \end{bmatrix}$$

The matrix of interest is the independent variable matrix in x:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \ldots\ldots & x_1^m \\ 1 & x_2 & x_2^2 \ldots\ldots & x_2^m \\ . & & & . \\ . & & & . \\ 1 & x_n & x_n^2 \ldots\ldots & x_n^m \end{bmatrix}$$

You may ask where the "1" in the first column of the matrix came from. Note that there is no term in x shown with β_0. However, it is there. It is x_n^0; any term raised to the zeroth power evaluates to 1, $\beta_0 \times 1 = \beta_0$. Trivial, for sure, but murder on your math if you forget it.

Now take a look back at the set of n equations at the top of the page. By analogy of form, the matrix equivalent of that set of equations is:

$$\mathbf{y} = \boldsymbol{\beta} \cdot \mathbf{X} + \mathbf{e} \qquad (3)$$

The least squares criterion says that the sum of the squares of the errors is to be minimized. Thus, we want to minimize the squared length of the error vector.

$$\begin{bmatrix} e_1 & e_2 & \dots & e_n \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} = \mathbf{e'} \cdot \mathbf{e} = |\mathbf{e}| = \text{a minimum} \qquad (4)$$

From equation 3, $\mathbf{e} = \mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta}$ and $\mathbf{e'} = (\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta})'$ giving us:

$$|\mathbf{e}| = (\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta}) \cdot (\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta})' \qquad (5)$$

The vector $\boldsymbol{\beta}$ represents the true values of the coefficients. We do not know these values; we are going to estimate them, and choose values which render the squared vector $|\mathbf{e}|$ a minimum. We will rename the vector $\boldsymbol{\beta}$ to $\mathbf{b}$, to reflect the fact that it is a vector of estimates, not true values. We can minimize the squared vector $|\mathbf{e}|$ by trial and error, but the mathematicians have done it already using calculus. The result is:

$$\mathbf{b} = (\mathbf{X'} \cdot \mathbf{X})^{-1} \cdot (\mathbf{X'} \cdot \mathbf{y}) \qquad (6)$$

And that's it. This is the general solution for the non-weighted regression case. Note that a solution exists if and only if the inverse matrix $(\mathbf{X'} \cdot \mathbf{X})^{-1}$ exists. The next step is to incorporate the weights.

We introduce the weight w of each y value in a diagonal weight matrix:

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & 0 \dots\dots & 0 \\ 0 & w_2 & 0 \dots\dots & 0 \\ . & & & . \\ . & & & . \\ 0 & 0 & 0 \dots\dots & w_n \end{bmatrix}$$

As in the previously covered weighted first order case, the weight is defined as the reciprocal of the variance scaled so that the sum of the weights is equal to the number of points. One now multiplies the error vector by the weights matrix, which effects the weighting:

$$\mathbf{W} \cdot \mathbf{e} = \begin{bmatrix} w_1 & 0 & 0 \dots\dots & 0 \\ 0 & w_2 & 0 \dots\dots & 0 \\ . & & & . \\ . & & & . \\ 0 & 0 & 0 \dots\dots & w_n \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} = \begin{bmatrix} w_1 \cdot e_1 \\ w_2 \cdot e_2 \\ . \\ . \\ w_n \cdot e_n \end{bmatrix}$$

Now follow this closely, referring back to the previous equations to see the analogy. As in equation (4), we now want to minimize the square length $|\mathbf{W}\cdot\mathbf{e}|$ of the weighted error vector, $\mathbf{W}\cdot\mathbf{e}$:

$$|\mathbf{W}\cdot\mathbf{e}| = (\mathbf{W}\cdot\mathbf{e})(\mathbf{W}\cdot\mathbf{e})' = (\mathbf{W}\cdot\mathbf{y} - \mathbf{W}\cdot\mathbf{X}\cdot\mathbf{b})\cdot(\mathbf{W}\cdot\mathbf{y} - \mathbf{W}\cdot\mathbf{X}\cdot\mathbf{b})'$$

We can clarify the analogy a bit by renaming:

$$\mathbf{W}\cdot\mathbf{y} = \mathbf{z} \quad \text{(it is a vector)}$$

$$\mathbf{W}\cdot\mathbf{X} = \mathbf{Z} \quad \text{(it is a matrix)}$$

and then $$|\mathbf{W}\cdot\mathbf{e}| = (\mathbf{z} - \mathbf{Z}\cdot\mathbf{b})(\mathbf{z} - \mathbf{Z}\cdot\mathbf{b})'$$

The form is identical to that of equation (5). Extending the analogy we obtain the solution in the manner of equation (6):

$$\mathbf{b} = (\mathbf{Z}'\cdot\mathbf{Z})^{-1} \cdot (\mathbf{Z}'\cdot\mathbf{z}) \qquad (6)$$

This may not seem like a lot of help, but let's now recast the solution in the notation involving the matrix $\mathbf{W}$:

$$\mathbf{b} = ((\mathbf{W}\cdot\mathbf{X})'(\mathbf{W}\cdot\mathbf{X}))^{-1}\cdot(\mathbf{W}\cdot\mathbf{X})'(\mathbf{W}\cdot\mathbf{y})$$

The basic properties of matrix multiplication come into play, now. Recall from Section 10: $\mathbf{A}\cdot\mathbf{B} \neq \mathbf{B}\cdot\mathbf{A}$, $(\mathbf{A}\cdot\mathbf{B})\cdot\mathbf{C} = \mathbf{A}\cdot(\mathbf{B}\cdot\mathbf{C})$, and $(\mathbf{A}\cdot\mathbf{B})' = \mathbf{B}'\cdot\mathbf{A}'$. Using these rules, we can rearrange our expanded solution this way:

$$\mathbf{b} = (\mathbf{X}'\cdot\mathbf{W}'\cdot\mathbf{W}\cdot\mathbf{X})^{-1}\cdot(\mathbf{X}'\cdot\mathbf{W}'\cdot\mathbf{W}\cdot\mathbf{y})$$

Now we have our solution expressed in terms of our original matrices, and it should be straightforward to set up in LOTUS. But, there is one simplification which will make the LOTUS solution easier. A diagonal matrix multiplied by its transpose yields a diagonal matrix with the original terms squared. You can set up a simple 2x2 example and try it. Applying this tidbit of insight to our solution, let's define:

$$\mathbf{U} = (\mathbf{W}'\cdot\mathbf{W}) = \begin{bmatrix} w_1^2 & 0 & 0 \ldots\ldots & 0 \\ 0 & w_2^2 & 0 \ldots\ldots & 0 \\ . & & & . \\ . & & & . \\ 0 & 0 & 0 \ldots\ldots & w_n^2 \end{bmatrix}$$

It is just as easy in LOTUS to set up the weights matrix with the squared weights as it is with the raw weights.

The final form of the matrix solution is then:

$$\mathbf{b} = (\mathbf{X}'\cdot\mathbf{U}\cdot\mathbf{X})^{-1}\cdot(\mathbf{X}'\cdot\mathbf{U}\cdot\mathbf{y})$$

You can calculate confidence limits by continuing with operations involving these matrices, plus a few others not too difficult to set up. You can define limits for each of the regression parameters, and for the predicted y given x. The interested reader is referred to chapter 13 of the excellent textbook on chemometrics by D.L. Massart et.al. (3).

12. The LOTUS 1-2-3 Spreadsheet for Weighted Parabolic Regression

Words are a poor substitute for an example when it comes to laying out a spreadsheet. Table 3 shows the look of the layout for the following exercise. The next few pages will guide you through the mechanics of creating it, and finding your answer. The text is set up in the format of instructions on the left, and highly illuminating commentary on the right. The example demo should require about 10 minutes if you're familiar with LOTUS, and about a half-hour if you're a novice. The data in the demo is real; it's from one of the Colourimetry Unit workstations. We will derive the second order weighted calibration curve.

Enter the headings "1", "x", and "x^2" in cells A1, B1, C1 respectively. Enter the concentration of each standard in column B, beginning in row 3. Enter 1 in column A for each entry in column B. Enter the formula +B3^2 in cell C3 and copy it to the remaining cells in the column, to the bottom of the range of the standards.

These columns *together* make up the independent variable matrix **X**. x is the concentration of our standards, and this is the only column of x-values that we enter. The 1 in column A is for the x^0 coefficient (see Section 11). The formula in column C calculates the square of each x-value for us.

Enter the heading "Y" in cell E1. Enter the responses in column E, in the row of the corresponding standard.

Column D is left blank for clarity. Column E holds the dependent variable vector, **y**. The y-values are the measured responses from the standards.

Head columns G,H, and I as "s", "1/s^2", and "w" respectively. Enter the standard deviation (or standard error when appropriate) associated with each standard in the corresponding row of column G. Enter the formula 1/G3^2 into cell H3, and copy it to the bottom of the range. Enter the formula @AVG(H3..H9) into cell H11. Enter the formula +H3/H$11 into cell I3, and copy it to the bottom of the range.

This section calculates the weight assigned to each *y*-value. The columns are the standard deviation (or standard error when we have several readings at each *x*), the reciprocal variance (weights), and the weighting factor. Cell H11 calculates the average reciprocal variance used in scaling the weights. The formula in column I (watch the $ sign) completes the calculation of the scaled weighting factors. Check your work by entering @SUM(I3..I9) in cell I11; the result should equal the number of data points.

Install the heading "U: weighting factors squared matrix" in cell K1. Enter a zero in cell K3, then copy cell K3 down to row 9, and right to column Q. Replace the zero in cell K3 with the formula +I3^2, the zero in cell L4 with +I4^2, the zero in cell M5 with +I5^2, and so on down the diagonal to cell Q9.

Here we create the diagonal matrix, **U**, of squared weighting factors. This matrix will extend for as many columns to the right as there are rows of data. In this example, fill 7 rows by 7 columns with zeroes. Then replace the diagonal elements only with a formula generating the square of the corresponding weighting factor from column I.

Users of LOTUS 1-2-3 v3.0 skip to the next instruction. Press /RV. Block cells C3..C9 as the range to copy from. Indicate cell C3 as the range to copy to, and press ENTER.

LOTUS 1-2-3 v2.0 and v2.01 try to transpose the formulae in column C, rather than the values, and get things screwed up. This step converts these formula cells to data cells, and all becomes well.

Enter the heading "X'" in cell A11. Press /RT. Block out cells A3..C9 as the range to transpose from and press ENTER. Position the cursor in cell A12 to mark the range to transpose to, and press ENTER.

This is **X'**, the *transpose* of the independent variable matrix. You can compare the position of the values in the transpose with those in the original matrix, to get a feel for how the transposition operator works.

Enter the heading "X'.U" in cell A16. Press /DMM. Block out the X' matrix, cells A17..G19, as the first matrix to multiply. Block out the U matrix, cells K3..Q9 as the second matrix to multiply. Position the cursor on cell A17 to mark the output range and press ENTER.

We multiply the transpose matrix of the independent variable by the weighting factors squared matrix. The order matters, so block the matrices in the order specified. Cell A17 marks the upper left corner of the output range of three rows by seven columns.

Enter the heading "X'.U.X" in cell A21. Press /DMM. Block out our just completed X'·U matrix as the first matrix to be multiplied. Block out the independent variable matrix A3..C9 as the second matrix to be multiplied. Indicate cell A22 as the output range and press ENTER.

X'·U figures in both of the two terms required to arrive at an answer. See the final expression in Section 11. This step continues towards formation of the first of those terms.

Enter the heading "inv(X'.U.X)" in cell A26. Press /DMI. Block out the newly formed product matrix A22..C24 as the range to invert. Indicate cell A27 as the output range, and press ENTER.

We have to invert the matrix we just created. The values in the resulting matrix will not seem to bear much relationship to the original matrix. If you want to prove the property of inverses for yourself, multiply the two matrices into an open area of the worksheet. You should get the identity matrix: a diagonal matrix of 1's.

Enter the heading "X'.U.Y" into column A31. Press /DMM. Block the X'·U matrix, A17..G19, as the first matrix to multiply. Highlight the column of *y*-values, cells E3..E9 as the second matrix. Indicate cell A32 and as the output range and press ENTER.

You've probably noticed that we have'nt involved the dependent variable, the responses, in the calculations as yet. Now is the time.

Enter the heading "b" in cell A36. Press /DMM and block out the "inv(X'.U.X)" matrix, cells A27..C29. Next block out the X'·U·y matrix, cells A32..A34. Indicate cell A37 as the output range and press ENTER. The resultant column vector is the answer set.

The final step is to multiply $(\mathbf{X'\cdot U\cdot X})^{-1}$ by $\mathbf{X'\cdot U\cdot y}$. The coefficients appear as a column vector in which the rows match the order of the columns in the independent variable matrix. In this case, they are β_0, β_1, and β_2 in the equation $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

QED.

It is not difficult to imagine this schema expanded for any combination of functions of the independent variable, or for multiple independent variables. Simply set up the independent variable columns in a block, so that they form a single matrix, as we did in the example. The size does not matter, save that you will have to move the other matrices to make room or that you may run out of memory. Enter the headings faithfully, and you should have no trouble keeping things straight. LOTUS 1-2-3 v.2.0-2.01 users remember to convert the data, in any columns of the independent variable matrix which are calculated from formulae, to values using the /RV function.

13. Summary

Linear regression is capable of determining a least-squares fit to curved lines as well as straight lines. The regression coefficients are constrained to linearity, not the dependent variable(s).

Three significant assumptions apply to ordinary linear regression:

1. - the independent variable values are known without error;

2. - the dependent variable errors are normally distributed;

3. - the standard deviation of the dependent variable values is the same for all.

When the first assumption is violated, you must use a technique such as the Maximum Liklihood Functional Relationship procedure presented in (4); the procedures presented in this report cannot treat significant variance in the independent variable. You can tame violations of the second assumption by applying a suitable transform, by applying robust statistics,

by substituting the mean of several readings for individual data points, or by suitably combining of all of these. Third assumption violations compel you to adopt a weighted regression procedure.

I have introduced you to the χ^2 test, which assesses the goodness of fit of your data to your chosen model. Applicable to all varieties of regression, this test is considered to be superior to evaluating the correlation coefficient in appraising the fit of the data to the curve.

In this report I have presented methods, in theory and with instructions and examples of LOTUS 1-2-3 solutions, for univariate and multivariate linear regression. The report covers both un-weighted and weighted solutions. One is again cautioned, though, that the methods set forth in this report all presume that the first assumption is valid: that the independent variable is free of significant error.

14. References

1. Miller, J.C. and Miller J.N.; *Statistics for Analytical Chemistry, Second Ed.*; John Wiley and Sons; New York, N.Y.; 1988

2. Freund, R.J. and Minton, P.D.; *Regression Methods: a Tool for Data Analysis*; Marcel Dekker, Inc; New York, N.Y.; 1979

3. Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y. and Kaufman, L.; *Chemometrics: a Textbook*; Elsevier Science Publishing Co. Inc.; New York, N.Y.; 1988

4. Rawlings, M.W.; *Standard Operating Procedure for Method Intercomparison*; MOE/LSB/WQS Internal Report; 1990

5. Analytical Methods Committee, Royal Society Of Chemistry; *Analyst*, 1988, **113**, 1469-1471

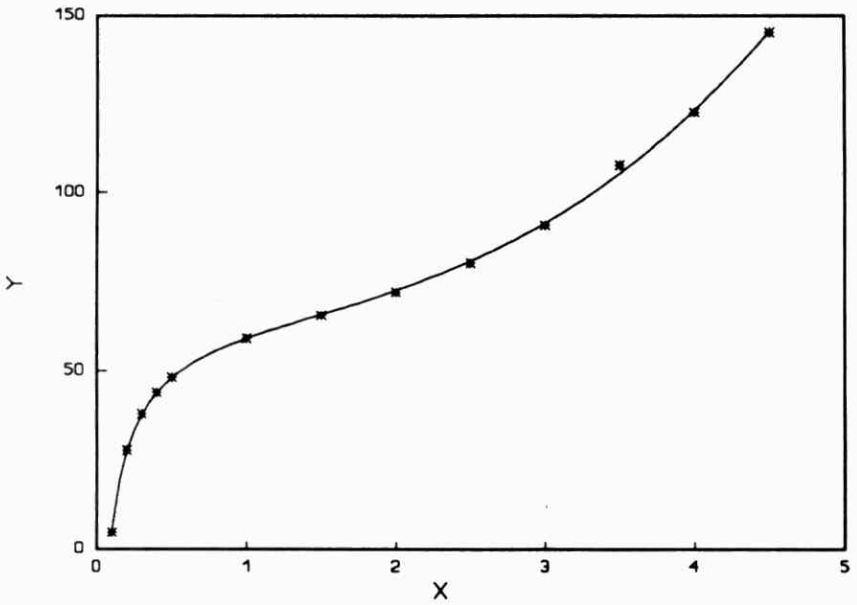
6. Whitney, D.R.; *Elements of Mathematical Statistics*; Holt, Rinehart and Winston; New York, N.Y.; 1961

Table 1: **LOTUS 1-2-3 Calculation of a Curvilinear Univariate Linear Regression**

	A	B	C	D
1	x	x^3	1/sqrt(x)	y
2	=======	=======	=======	=======
3	0.1	0.001	3.162278	4.8
4	0.2	0.008	2.236068	27.8
5	0.3	0.027	1.825742	37.9
6	0.4	0.064	1.581139	44
7	0.5	0.125	1.414214	48.2
8	1	1	1	59.2
9	1.5	3.375	0.816497	65.7
10	2	8	0.707107	72.1
11	2.5	15.625	0.632456	80.2
12	3	27	0.57735	90.8
13	3.5	42.875	0.534522	107.7
14	4	64	0.5	122.6
15	4.5	91.125	0.471405	145.3
16				
17		Regression Output:		
18	Constant			82.03073
19	Std Err of Y Est			0.888377
20	R Squared			0.999621
21	No. of Observations			13
22	Degrees of Freedom			9
23				
24	X Coefficient(s)	0.72807	0.788337	-24.3768
25	Std Err of Coef.	0.713838	0.027579	0.678922

The resulting equation for the data is:

$$y = 82.03073 + 0.72807x + 0.788337x^3 - 24.3768/\sqrt{x}$$



Data from Table 1 with Regression Curve

Table 2: **Lotus 1-2-3 Summations Solution for Weighted First Order Linear Regression**

	===A===	===B===	===C===	===D===	===E===	===F===	===G===	===H===	===I===
1	x	y	s	1/s^2	w	wixi	wiyi	wixiyi	wixi^2
2	========	========	========	========	========	========	========	========	========
3	0	0.009	0.001	1000000	5.535344	0	0.049818	0	0
4	2	0.158	0.004	62500	0.345959	0.691918	0.054662	0.109323	1.383836
5	4	0.301	0.01	10000	0.055353	0.221414	0.016661	0.066646	0.885655
6	6	0.472	0.013	5917.16	0.032754	0.196521	0.01546	0.092758	1.179127
7	8	0.577	0.017	3460.208	0.019153	0.153228	0.011052	0.088412	1.22582
8	10	0.739	0.022	2066.116	0.011437	0.114367	0.008452	0.084517	1.143666
9									
10			Avg.	1/s^2:	180657.2	Sums:	1.377447	0.156104	0.441656	5.818104
11									
12	Weighted centroids:			x:	0.229574	y:	0.026017		
13									
14		Coefficients:		B0:	0.009084	B1:	0.07376		

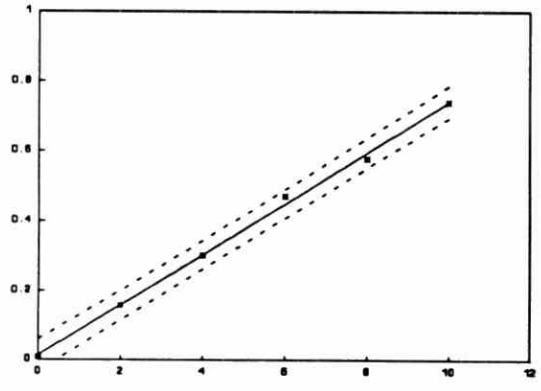
The resulting equation for the data is:

$$y = 0.009084 + 0.07376x$$

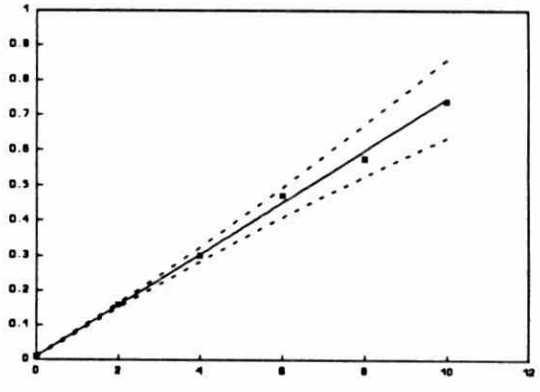
Ordinary linear regression yields the equation:

$$y = 0.013286 + 0.07254x$$

Looking at the equations, you cannot see a great deal of difference; this happens frequently. Calibration data, for example, is inherently precise. The significant difference between the weighted and un-weighted methods is in the confidence with which y-values may be predicted from the curve given an x-value. The figures below show the weighted and un-weighted lines and their respective confidence limits. Note the distinct differences in the shape and proximity to the regression line in the two cases. Clearly the weighted case better recognizes the variability in the variance of the y values.

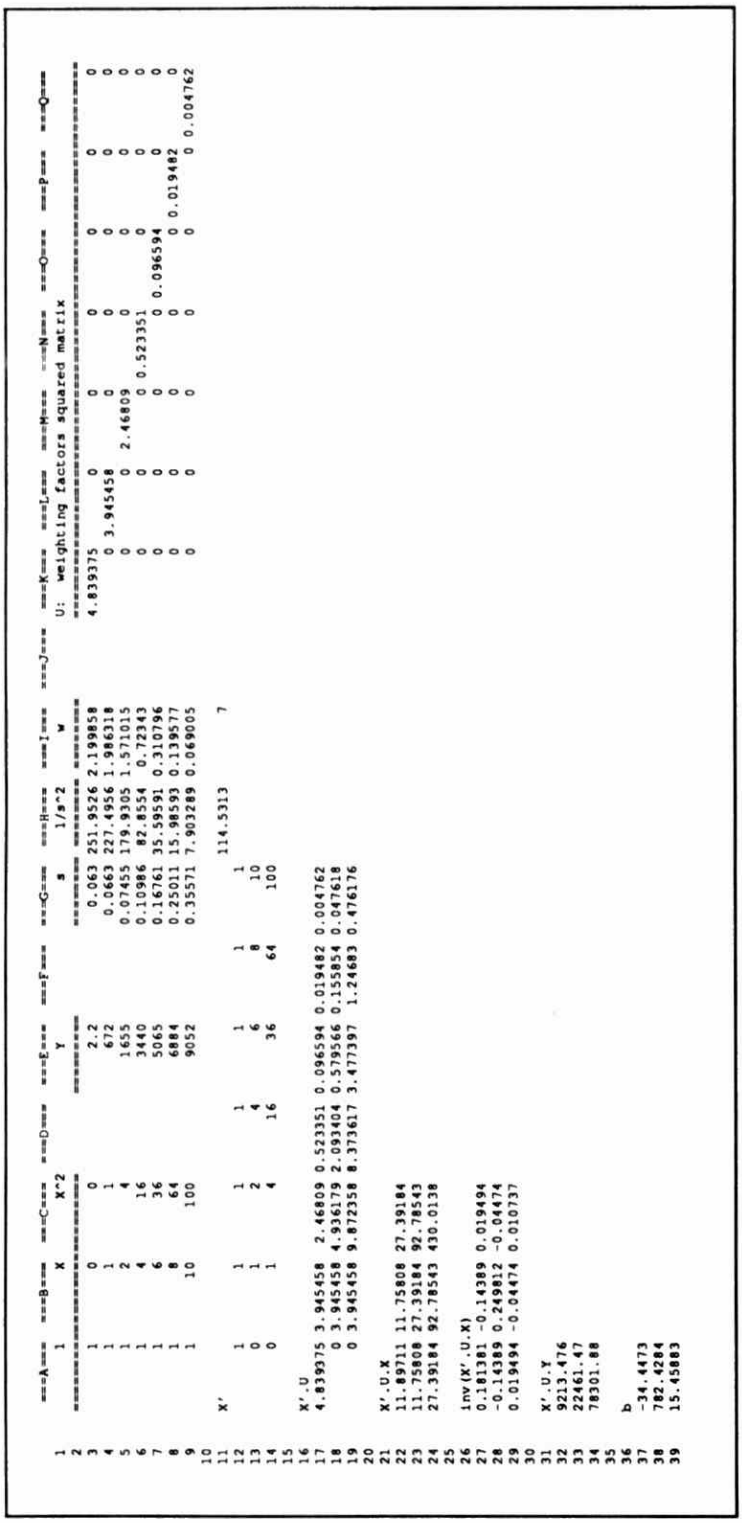


Confidence Band with Ordinary Regression



Confidence Band with Weighted Regression

Table 3: **LOTUS 1-2-3 Matrix Solution for Weighted Second Order Linear Regression**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	1	X	X^2		Y		s	1/s^2	w		U: weighting factors squared matrix						
2																	
3		1	0	0	2.2		0.063	251.9526	2.199858		4.839375	0	0	0	0	0	0
4		1	1	1	672		0.0663	227.4956	1.986318		0	3.945458	0	0	0	0	0
5		1	2	4	1655		0.07455	179.9305	1.571015		0	0	2.46809	0	0	0	0
6		1	4	16	3440		0.10986	82.85554	0.72343		0	0	0	0.523351	0	0	0
7		1	6	36	5065		0.16761	35.59591	0.310796		0	0	0	0	0.096594	0	0
8		1	8	64	6884		0.25011	15.98593	0.139577		0	0	0	0	0	0.019482	0
9		1	10	100	9052		0.35571	7.903289	0.069005		0	0	0	0	0	0	0.004762
10																	
11	X'							114.5313		7							
12		1	1	1	1	1	1										
13		0	1	2	4	6	8										
14		0	1	4	16	36	64	100									
15																	
16	X'.U																
17	4.839375	3.945458	2.46809	0.523351	0.096594	0.019482	0.004762										
18	0	3.945458	4.936179	2.093404	0.579566	0.155854	0.047618										
19	0	3.945458	9.872358	8.373617	3.477397	1.24683	0.476176										
20																	
21	X'.U.X																
22	11.89711	11.75808	27.39184														
23	11.75808	27.39184	92.78543														
24	27.39184	92.78543	430.0138														
25																	
26	inv(X'.U.X)																
27	0.181381	-0.14389	0.019494														
28	-0.14389	0.249812	-0.04474														
29	0.019494	-0.04474	0.010737														
30																	
31	X'.U.Y																
32	9213.476																
33	22461.47																
34	78301.88																
35																	
36	b																
37	-34.4473																
38	782.4284																
39	15.45883																

Use the raw data and this example as a guide in laying out the spreadsheet and carrying out the calculations. The raw data is:

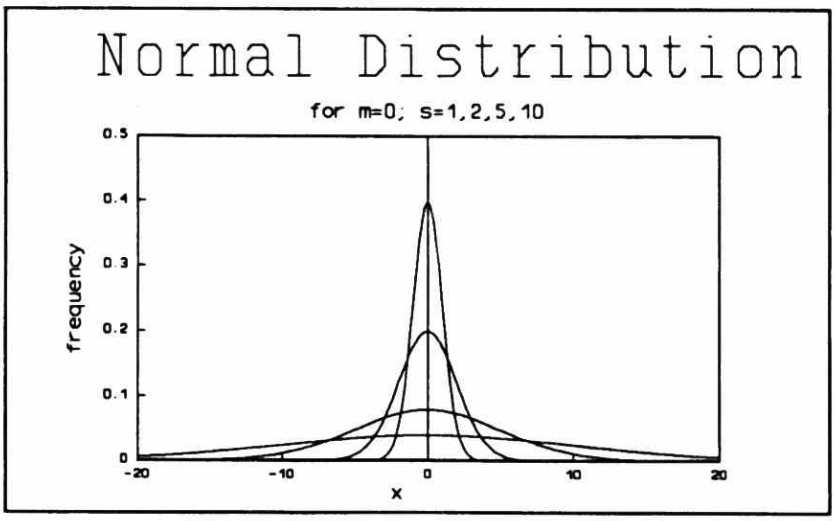
X	Y	s
0	2.2	0.063
1	672	0.0663
2	1655	0.07455
4	3440	0.10986
6	5065	0.16761
8	6884	0.25011
10	9052	0.35571

Generate the remaining columns by entering the formulae and executing the operations described in the text.

The resulting equation for the data is:

$$y = -34.4473 + 782.4284x + 15.45883x^2$$

APPENDIX 1:
The Normal Distribution



Suppose we made a large number of unbiased replicate measurements for which we know the true value. Because of random errors in the measurement process, we will end up with a large number of different values, many occurring more than once, grouped around the expected value. Then we examine the errors by subtracting the true value from each reading. The errors remain grouped in the same pattern as the readings, but around zero. If we now plot the errors on the x-axis, versus the number of times each value occurs on the y-axis, we obtain the distribution of the errors. If we divide the number of times a value occurs by the total number of readings, the graph becomes the frequency distribution of the errors.

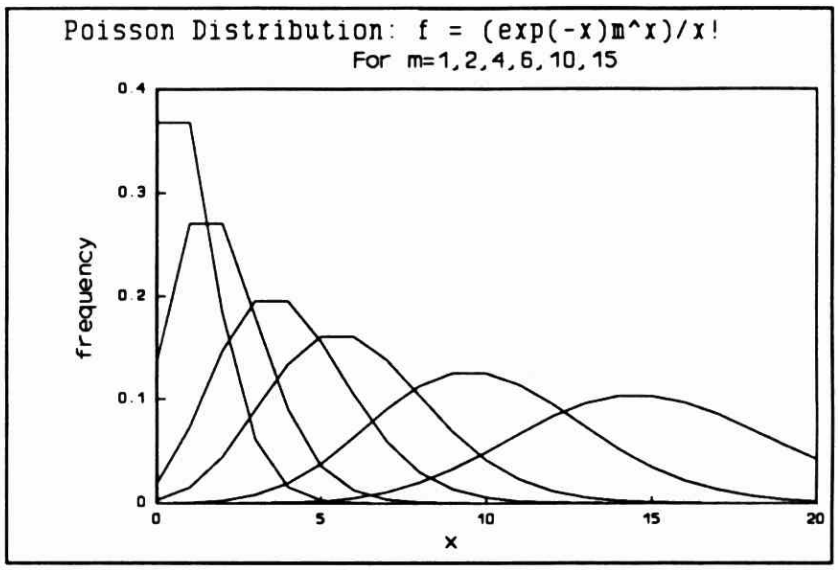
Workers in all branches of the physical, social, and natural sciences , beginning with the earliest "philosophers", have seen this shape of distribution occur repeatedly. We encounter it in chemical measurements, yields of grain from various fields, lengths of fish in various lakes, heights of children in various populations, proportion of spotted beans in various gardens. As the sciences developed, various workers collectively characterized this distribution, placed it on a sound theoretical foundation, then used it to develop a wide variety of descriptive and analytical tools we today know as statistics. This fundamental distribution of measurement data is named the Normal Distribution.

The distribution is symmetric about its central axis, the arithmetic mean. The broadness of the distribution is quantified by the standard deviation. The figure illustrates the effect of increasing standard deviation on a distribution of constant mean. As you might guess from the way the curve is squashed as standard deviation rises, the total area under the normal curve is always 1. You can find the probability that any given reading will occur by taking the area under the curve covered by the interval of the reading. (For example, for the reading 2 in the series 1,2,3 ...., the interval covered is any value between 1.5 and 2.5.)

This is not the only distribution known to describe measurement processes. It is by far the most common and useful in summarizing a data set, even if it is only a first approximation. It has the advantage of being a universal common ground, facilitating communication among scientists of all disciplines.

APPENDIX 2:
The Poisson Distribution

A second useful distribution of data, often found in sciences involving the counting of random events, is the Poisson distribution. This distribution represents the probability that an event will occur within a given time period, such as in radioactive decay.



The curves at right illustrate the shape of the distribution and clearly show the effect of changing the mean. In fact the distribution is discontinuous; we should show it as discrete bars at integer values of count (x-axis), with height equal to the probability at that point. That is because you do not have fractional events.

The mean is the arithmetic mean, as with the normal distribution. The standard deviation, however, is the square root of the mean. Since the distribution is asymmetric, one should not apply the same criteria or hold the same expectations of behaviour for Poisson distributed data as for normally distributed data. Poisson statistics demand separate treatment.

Are there statistical distributions other than the Normal and the Poisson with which we are in common contact? Of course, but we seldom make direct measurements involving them. The binomial distribution, for instance, accurately describes the statistics of flipping a coin, or tossing dice. If we delve into the quantum-mechanical description of particles with half-integer spin, such as electrons or protons, we would invoke Bose-Einstein statistics. Each of these has its own mathematical peculiarities which must be understood and applied when the distribution underlying the data fits the particular model. Do bear in mind this property of means, however; that the means of readings taken from any distribution will themselves be normally distributed. When you're having difficulty with a badly behaved or uncertain distribution of data, take several readings at each point and determine the means. You can then correctly use the well known Normal Distribution statistics on the means for your evaluations. See any good statistics text for guidance and examples.

(8114)
TA/340/R37/MOE